IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wakamori Chiaki
3/16/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

Summary of methodologies

- Collecting the Data
- Data Wrangling
- Exploratory Data Analysis using SQL, Pandas and Matplotlib
- Interactive Visual Analytics and Dashboards
- Predictive Analysis Overview

Summary of all results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

- Project background and context

One of the reasons SpaceX is able to make numerous launches possible is that the rockets are relatively inexpensive to launch.

The SpaceX website lists the cost of launching a Falcon 9 rocket at $62 million, while other providers cost more than $165 million each. Therefore, much of the savings is due to SpaceX's ability to reuse the first stage. Thus, if we can determine if the first stage will land, we can determine the cost of the launch.

- Problems you want to find answers

To use data to predict whether SpaceX will succeed in landing the first stage of the rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Reset API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection

- How data sets were collected.

Collect data from SpaceX Reset API and Web Scraping from Wikipedia

SpaceX reset API

decode the response content as a Json

turn it into a Pandas dataframe

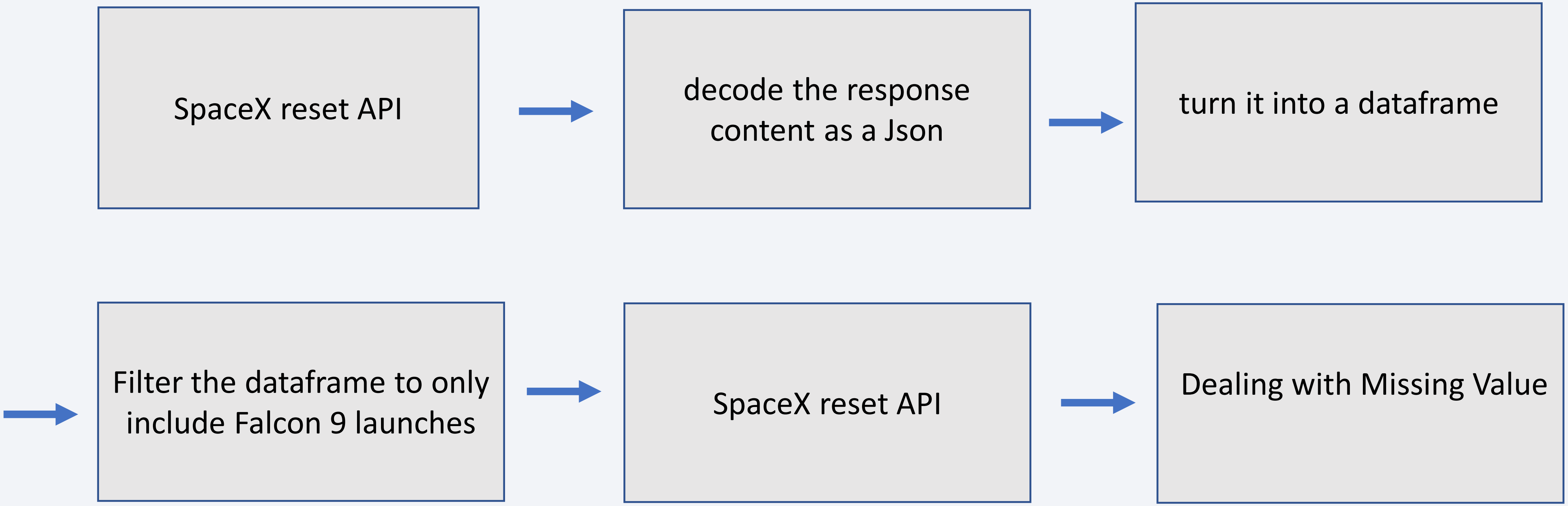
Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response

Extract column/variable names from HTML table headers using beautifulsoup

Create a data frame by parsing the launch HTML tables

Data Collection – SpaceX API

Collect data from SpaceX Reset API



https://github.com/waka1234/IBM_WatsonStudio/blob/master/Data%20Collection%20API.ipynb

Data Collection - Scraping

Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia

Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response → Extract column/variable names from HTML table headers using beautifulsoup → Create a data frame by parsing the launch HTML tables

https://github.com/waka1234/IBM_WatsonStudio/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models

Calculate the number of launches on each site → Calculate the number and occurrence of each orbit → Calculate the number and occurence of mission outcome per orbit type → Create a landing outcome label from Outcome column

https://github.com/waka1234/IBM_WatsonStudio/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

The function catplot is used to represent the following relationship in a diagram
Relationship between flight number and launch site
Relation between launch site and payload mass
Relationship between Success Rate and Orbit Type Relationship between Success Rate and Orbit Type
Plot relationship between flight number and orbit type
Relationship between payload and orbit type

https://github.com/waka1234/IBM_WatsonStudio/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite%20.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL

- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

https://github.com/waka1234/IBM_WatsonStudio/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map

Add a highlighted circle area with text label at specific coordinates using folium

Create and add folium.Circle and folium.Marker for each launch site on the site map

- Mark the success/failed launches for each site on the map

Create a new column in launch_sites dataframe called marker_color to store the marker colors based on the class value

For each launch result in spacex_df data frame, add a folium.Marker to marker_cluster

- Calculate the distances between a launch site to its proximities

Add Mouse Position to get the coordinate (Lat, Long) for a mouse over on the mapMark down a point on the closest coastline using MousePosition and calculate the distance between the coastline point and the launch site

https://github.com/waka1234/IBM_WatsonStudio/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite%20.ipynb

Build a Dashboard with Plotly Dash

Building a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time

- Add a Launch Site Dro-down Input Component
- Add a callback function to render success-pi-chart based on selected site dropdown
- Add a Range Slider to Select Payload
- Add a callback function to render the success-payload-chart scatter plot

https://github.com/waka1234/IBM_WatsonStudio/blob/master/space_dash_app.py

Predictive Analysis (Classification)

create a machine learning pipeline to predict if the first stage will land given the data

The models are trained and hyperparameters are selected using the function GridSearchCV

→

Create a logistic regression object

→

Create a support vector machine object then create a GridSearchCV object svm_cv, tree_cv, knn_cv

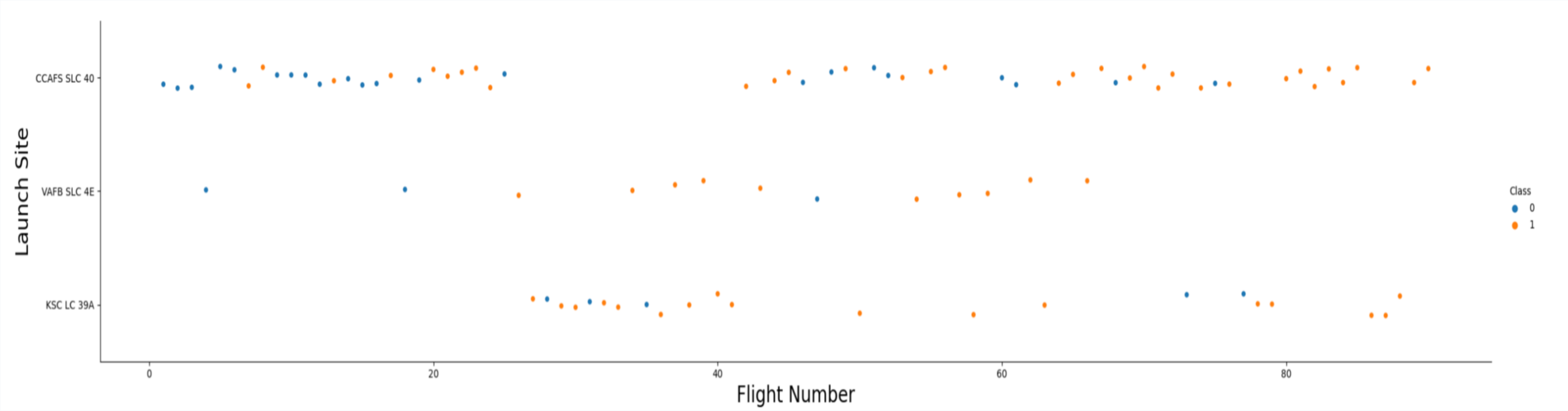
https://github.com/waka1234/IBM_WatsonStudio/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Section 2

Insights drawn from EDA

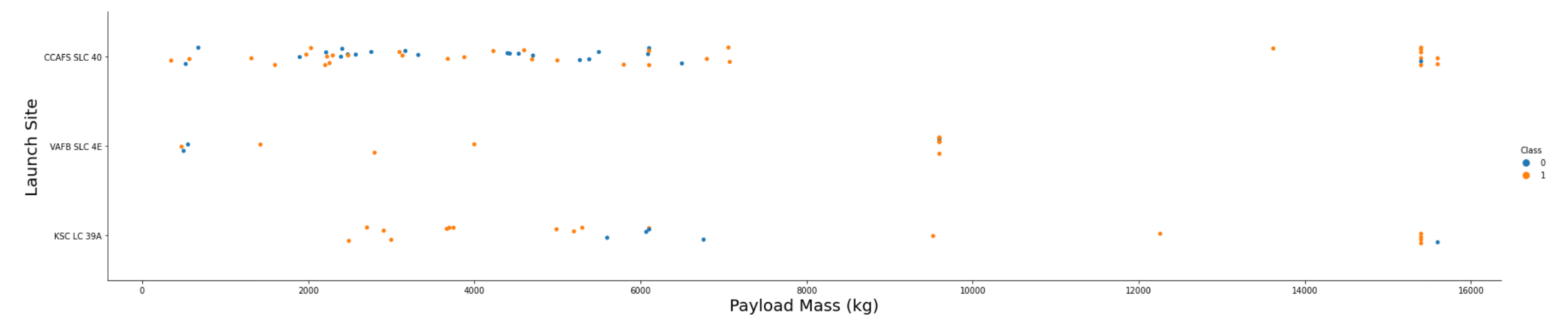
Flight Number vs. Launch Site

- VAFB SLC 4E and KSC LC 39A have high success rates
- CCAFS SLC 40 success rate increases Flight Number



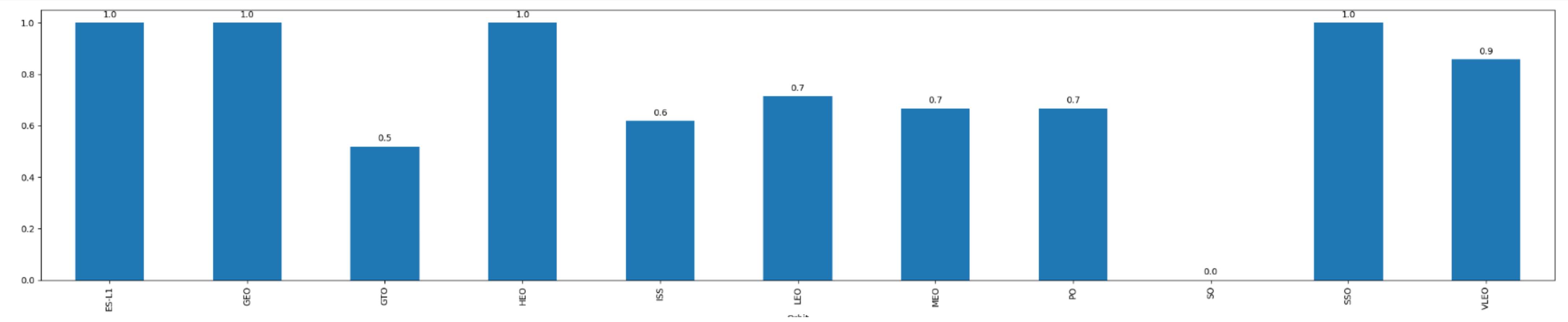
Payload vs. Launch Site

- Success rate increases when Payload Mass is over 8000 kg
- Failure rate is high for KSC LC 39A when Payload Mass is between 6500 and 7500 kg
- VAFB SLC 4E has a high success rate
- CCAFS SLC 40, the success/failure ratio does not change much for payloads under 8000 kg, but is higher for payloads over 12000 kg



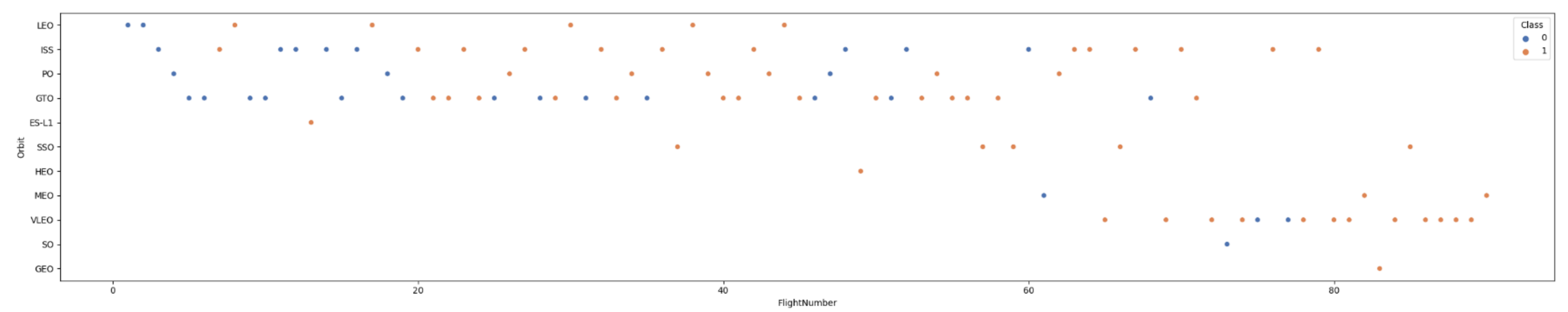
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have a 100% success rate
- All but GTO, ISS, and SO have high success rates
- SO has a success rate of 0%



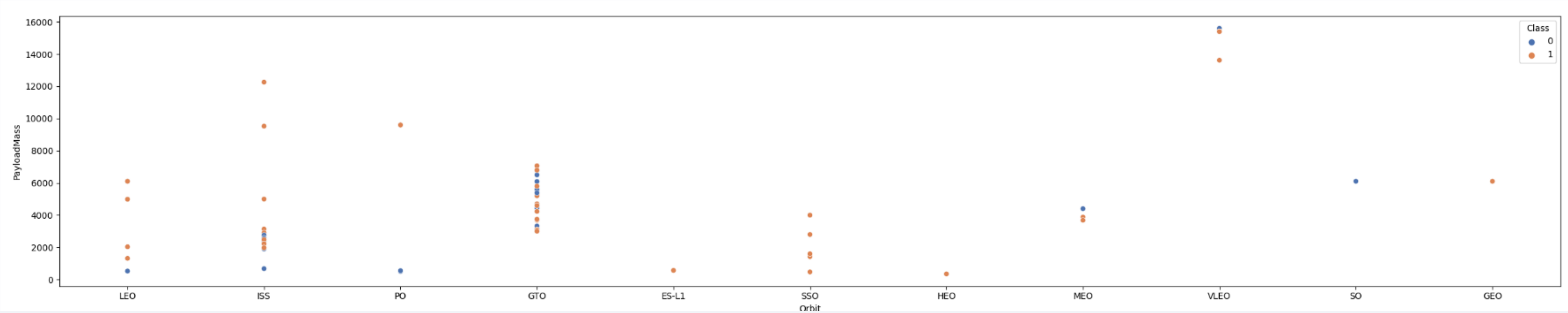
Flight Number vs. Orbit Type

- Success Rate of 100% has only one Orbit data in Flight Number
- Success rate tends to increase as Flight Number increases



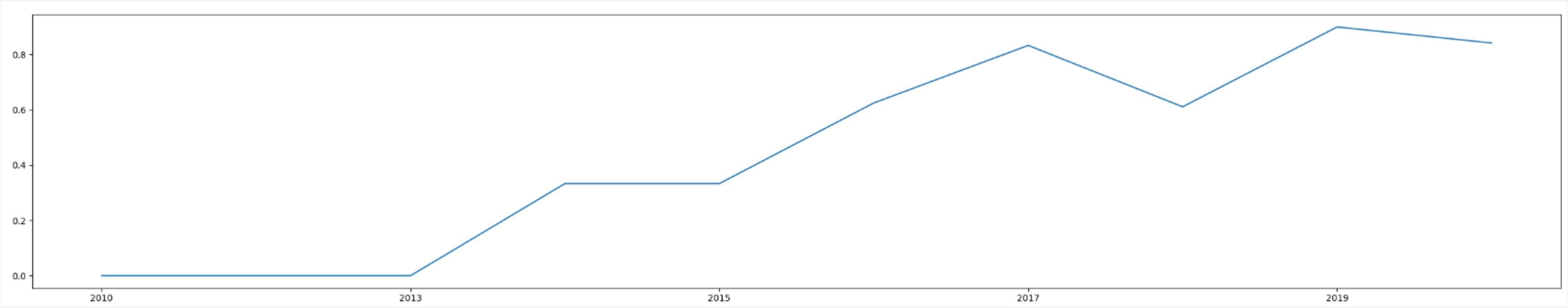
Payload vs. Orbit Type

- For LEO and ISS, the higher the Payload Mass, the higher the success rate
- For GTO, the higher the Payload Massn value, the lower the success rate



Launch Success Yearly Trend

- Since 2016, our success rate is over 60%
- Since 2019, the success rate has been over 80%



All Launch Site Names

Use DISTINCT to display unique launch site names for space missions

```
%sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

```
* ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897
  sqlite:///my_data1.db
Done.
```

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

Use WHERE and LIKE to display 5 records whose launch site begins with the string "CCA".

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
* ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
  sqlite:///my_data1.db
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Use WHERE clause to show the total onboard mass of the booster launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

```
 * ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od
   sqlite:///my_data1.db
Done.
```

1
45596

Average Payload Mass by F9 v1.1

Use WHERE and AVE to indicate the average payload mass carried by the booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

```
 * ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8l
   sqlite:///my_data1.db
Done.
```

1
2928

First Successful Ground Landing Date

Use Min function to list the date of the first successful landing on the ground pad

```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME='Success (ground pad)'
```

```
 * ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg
   sqlite:///my_data1.db
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Use WHERE List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTCOME='Success (drone ship)'
```

```
 * ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
   sqlite:///my_data1.db
Done.
```

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Using COUNT,WHERE clause, LIKE, list the names of successful boosters on drone ships with a payload mass between 4000 and 6000
- Use WHERE clause to list the total number of mission successes and failures

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTCOME='Success (drone ship)'
```

```
* ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
  sqlite:///my_data1.db
Done.
```

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

```
* ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:
  sqlite:///my_data1.db
Done.
```

1
101

Boosters Carried Maximum Payload

Use Max Function to state the version name of the booster with the maximum loading capacity

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536,
   sqlite:///my_data1.db
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, ¥
   LANDING__OUTCOME AS LANDING__OUTCOME, ¥
   BOOSTER_VERSION AS BOOSTER_VERSION, ¥
   LAUNCH_SITE AS LAUNCH_SITE ¥
   FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%'
```

```
 * ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databas
   sqlite:///my_data1.db
Done.
```

month_name	landing__outcome	booster_version	launch_site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql SELECT "DATE", COUNT(LANDING__OUTCOME) as COUNT FROM SPACEXTBL ¥
    WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20' AND LANDING__OUTCOME LIKE '%Success%' ¥
    GROUP BY "DATE" ¥
    ORDER BY COUNT(LANDING__OUTCOME) DESC
```

```
* ibm_db_sa://ngs23261:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.a
  sqlite:///my_data1.db
Done.
```

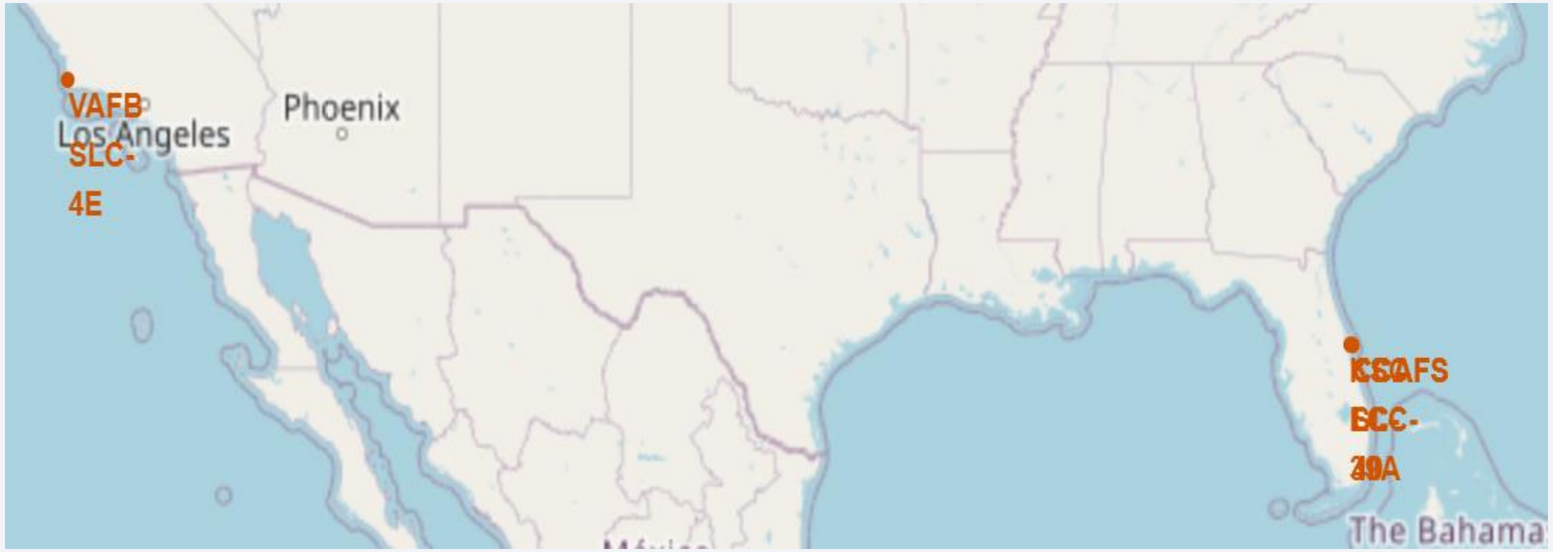
DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

Section 3

Launch Sites Proximities Analysis

<All launch site>

The following figure shows a map created from the SpaceX launch data, with the launch locations marked



<Success or failure of launch at each launch site>

The following image shows the location of the launch site

Red marks indicate a successful plane launch and green marks indicate a failed plane launch

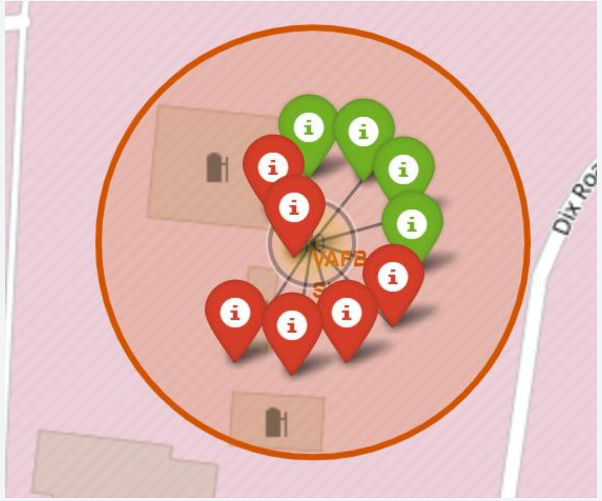


Fig.1 Launch site in VAFB SLC-4E

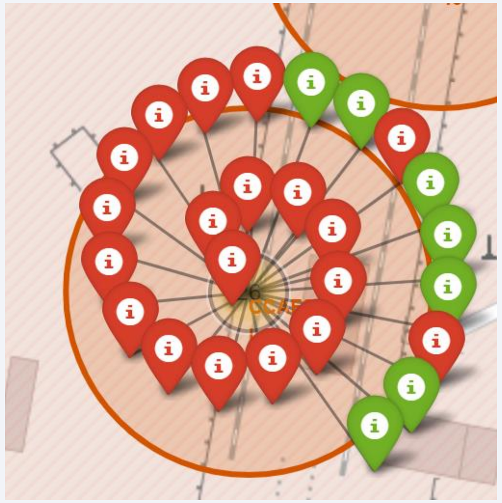


Fig.2 Launch site in CCAFS LC-40

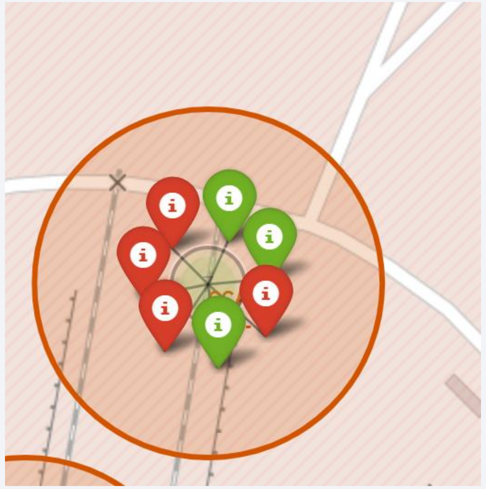


Fig.3 Launch site in CCAFS SLC-40

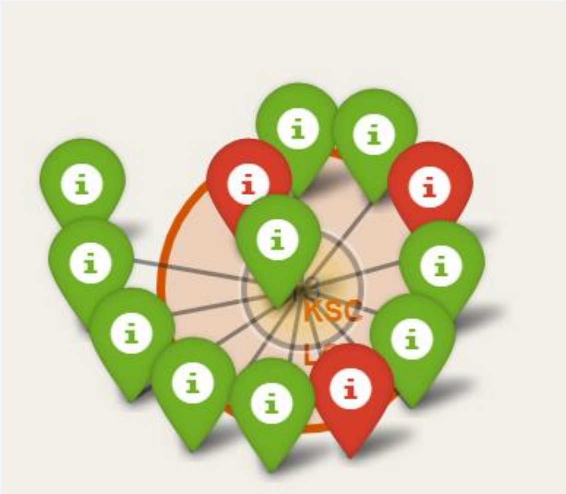
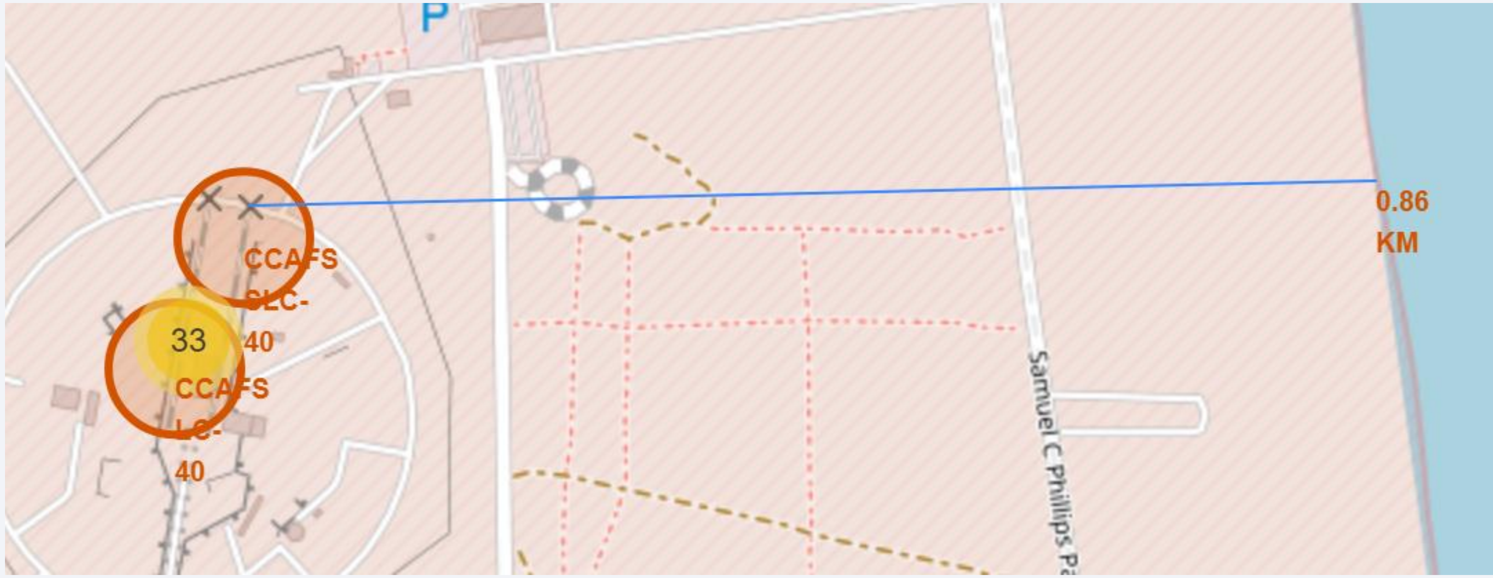


Fig.4 Launch site in KSC LC-9A

<Distance between a launch site to its proximities>

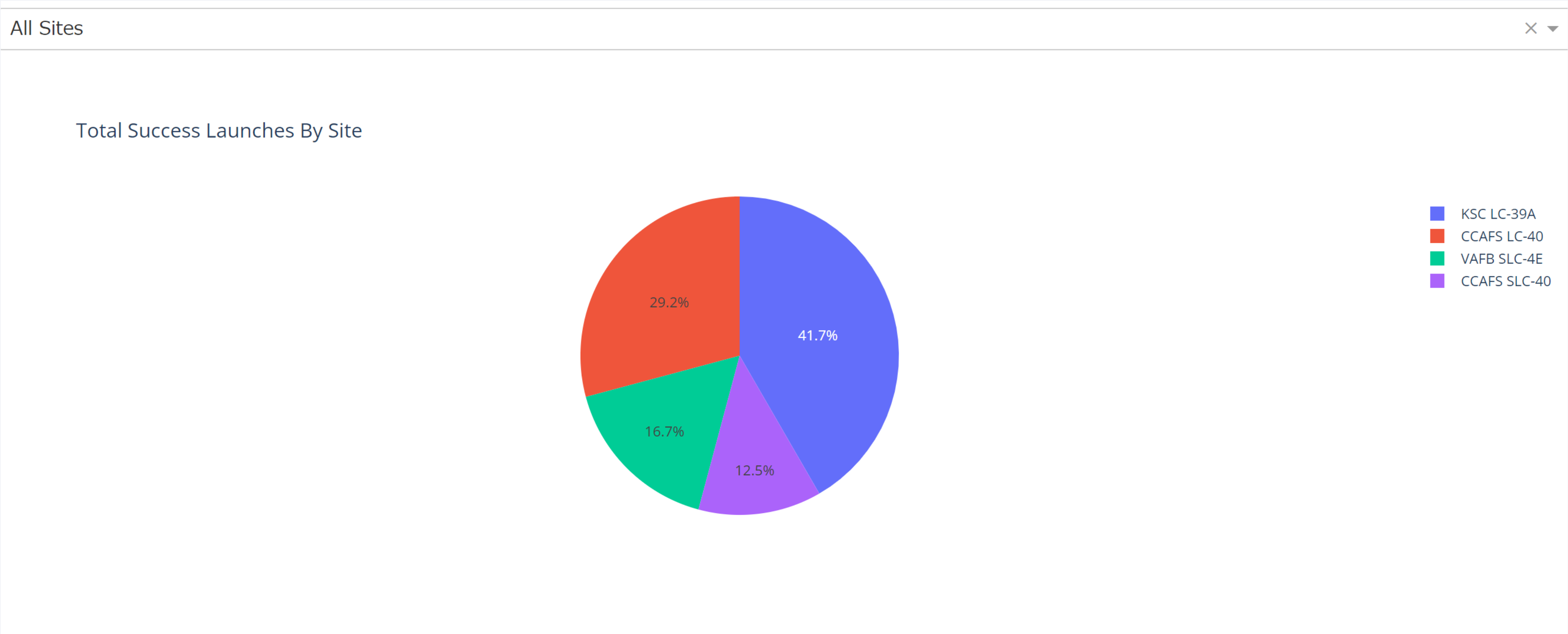
The figure below shows the result of marking the nearest shoreline point and calculating the distance from that shoreline point to the launch site



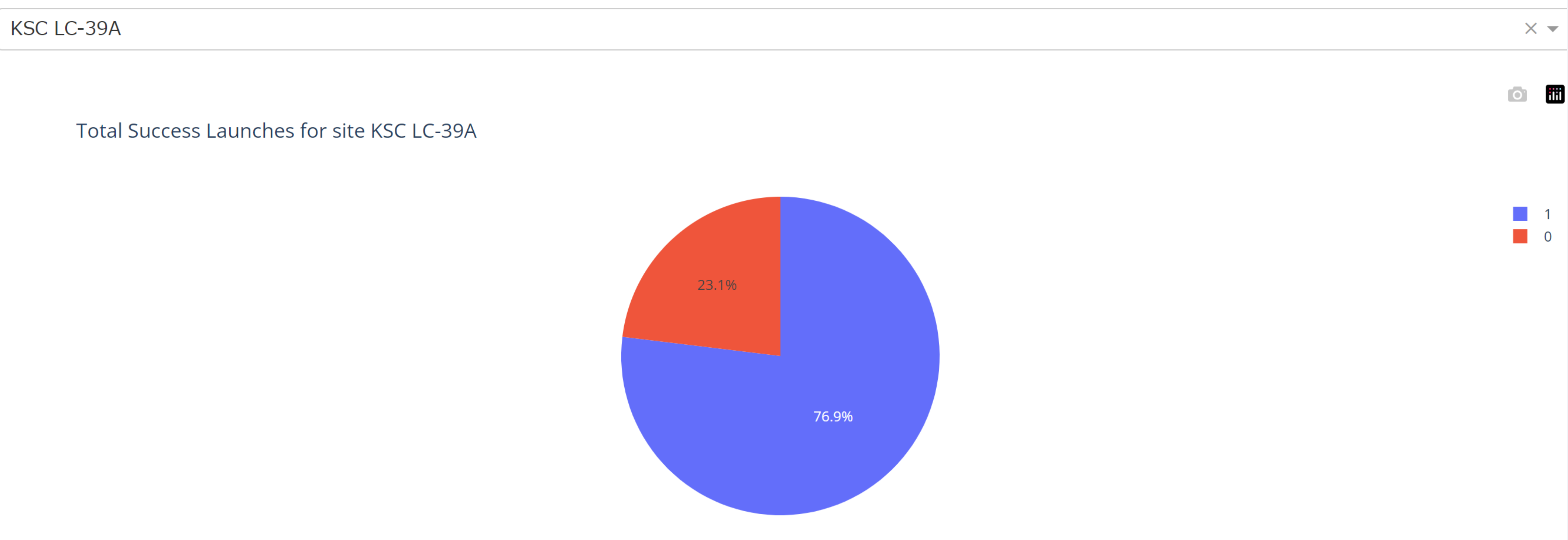
Section 4

Build a Dashboard with Plotly Dash

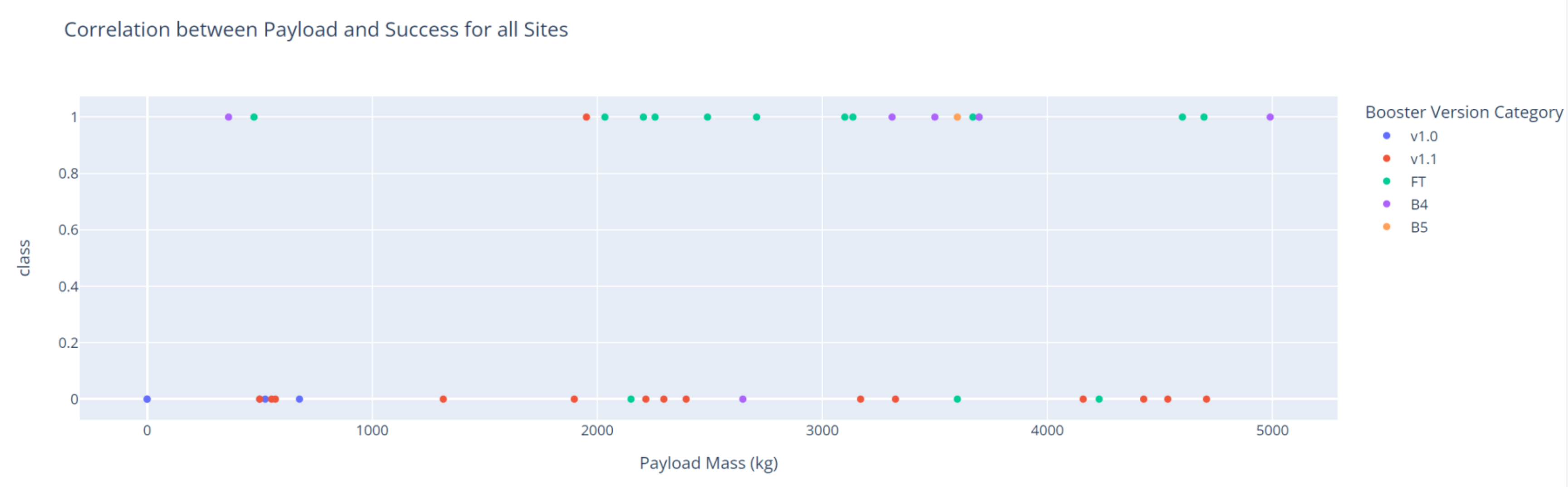
<Total Success Launches By Site>

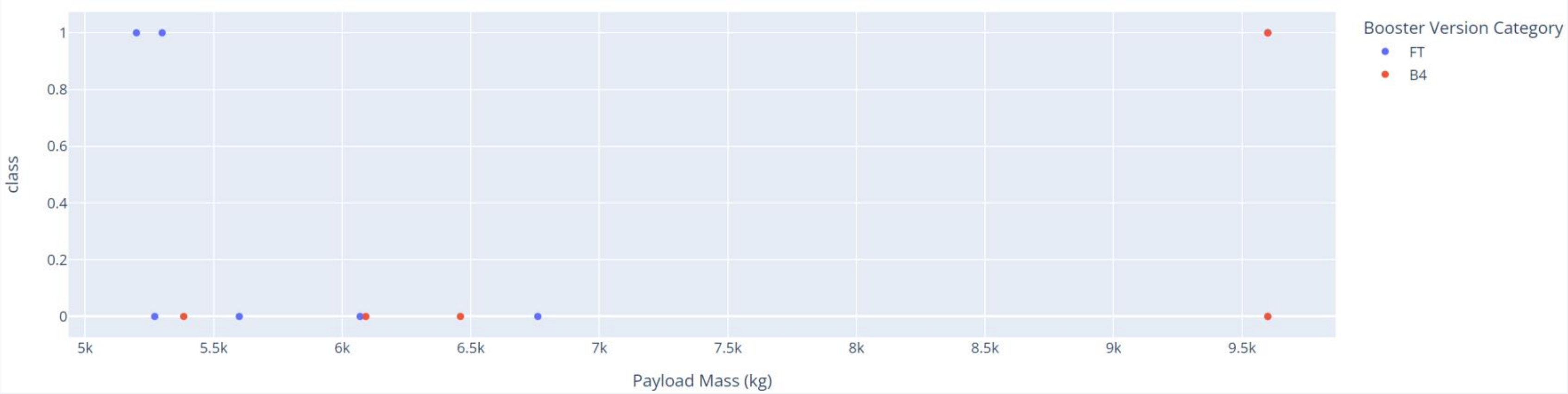


<Total Success Launches for site KSC LC-39A>



<Correlation between Payload and Success for all Sites>





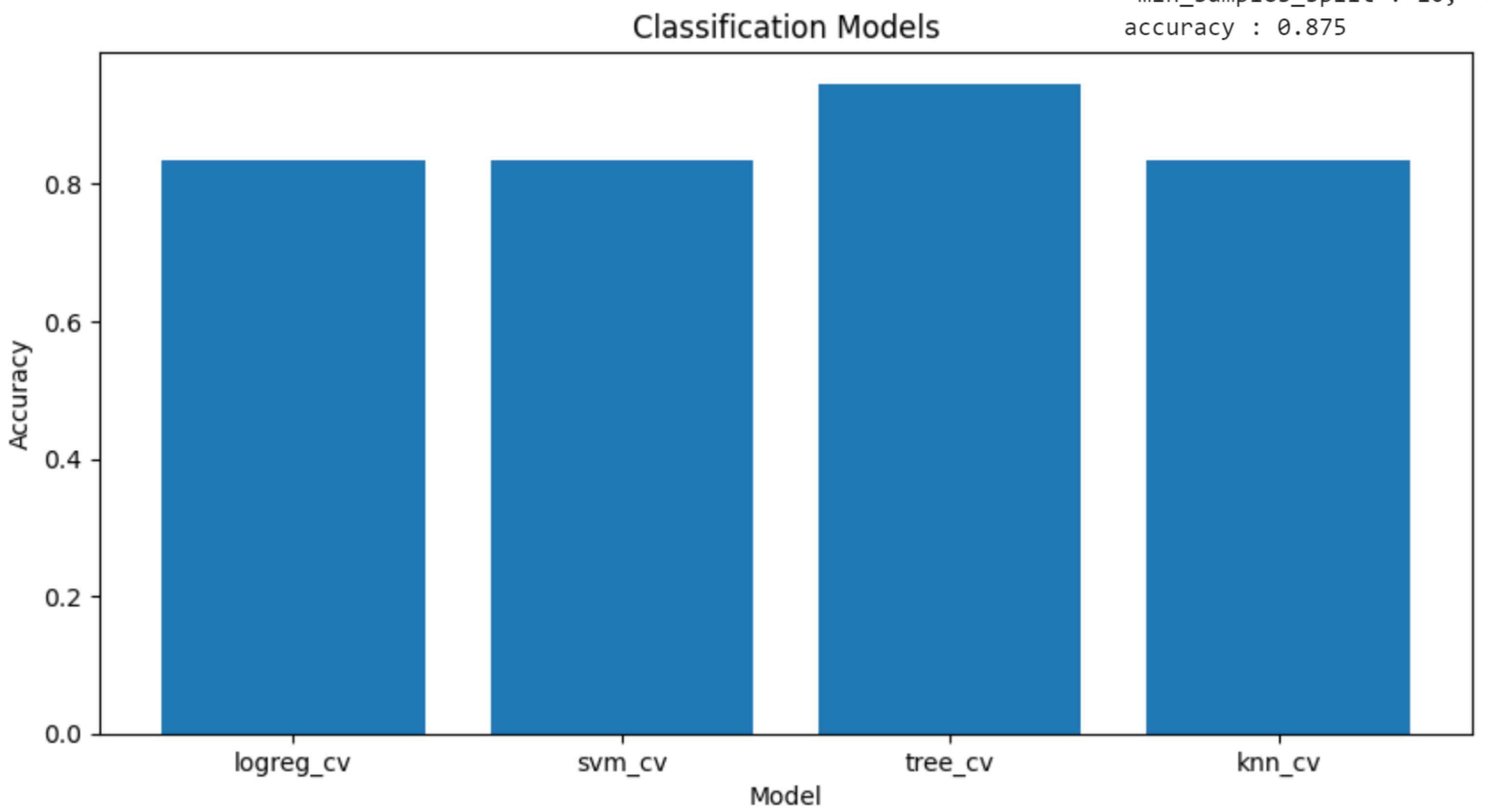
Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

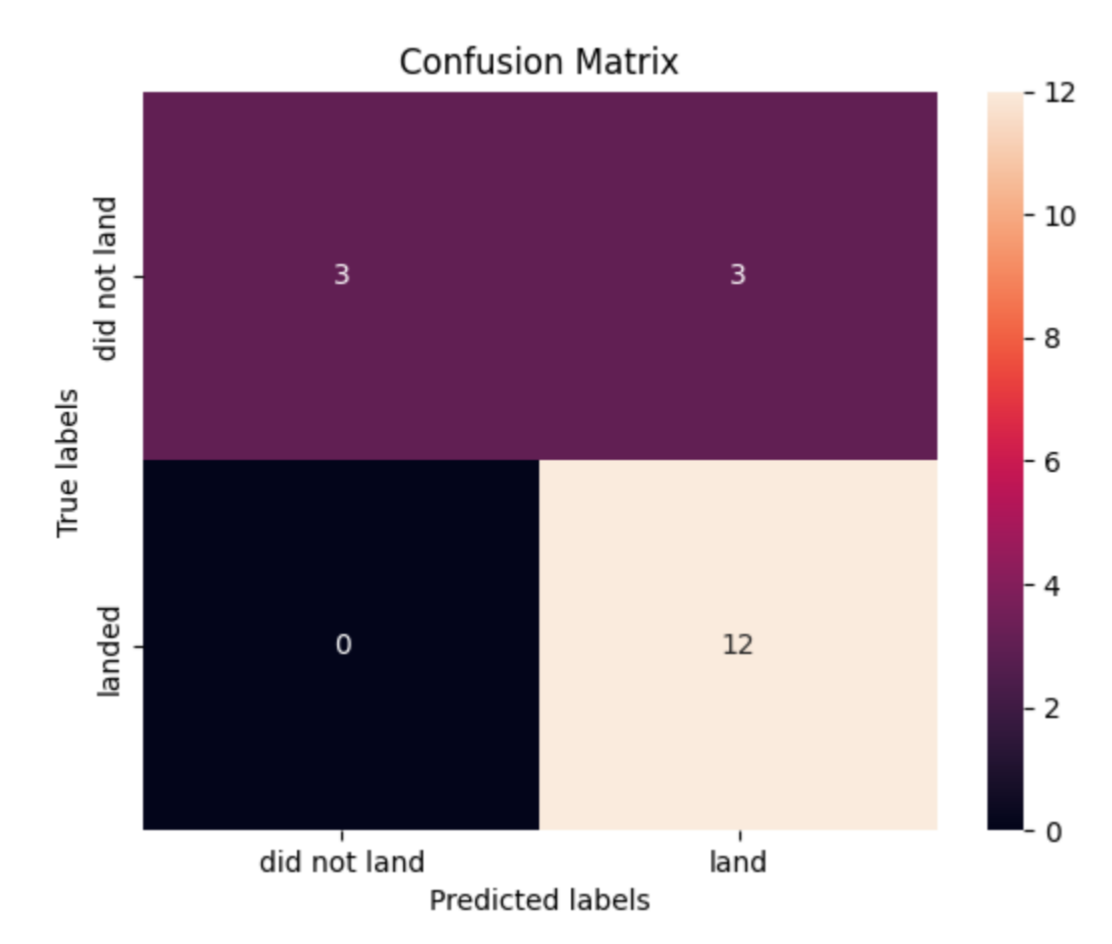
```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 18,
'min_samples_split': 10, 'splitter': 'best'}
accuracy : 0.875
```



- The most Accuracy is tree_cv and accuracy of 0.875
- logreg_cv, svm_cv and knn_cv accuracy of about 0.8

Confusion Matrix

The results of the Confunsion Matrix were as follows in all cases



- False positives have a value of 0
- The value of true positives and false negatives is 3
- The value of a true negative is 12

Conclusions

- Success rate tends to increase as flight number increases
- Success rate is more than 80% after 2019.
- It is better to launch when Payload Mass is more than 8000 kg
- When Payload is less than 6000 kg, the orbit with the highest success rate is SSO
- Success rate is higher when Launch site is in KSC LC-9A.
- When Payload Mass is less than 5000 kg, Booster Version Category should be FT
- The highest classification accuracy is tree_cv
- In the mixed matrix, false negatives are zero, so the confidence level is high

Thank you!